Synopsis

Watson-Crick paired B-form DNA is the genetic material in most of the biological systems. Integrity of DNA is of utmost importance for the normal functioning of any organism. Various environmental factors, chemicals and endogenous agents constantly challenge integrity of the genome resulting in mutagenesis. Over the past few decades multiple reports suggest that DNA can adopt alternative conformations other than the right handed double helix. Such structures occur within the context of B-DNA as sequence dependent structural variations and are facilitated by free energy derived from negative supercoiling, which may be generated during physiological processes like transcription, replication, etc. or binding of proteins. Multiple groups have shown that these structures render fragility to the genome owing to single-strandedness (presence of unpaired bases). This conformational polymorphism of the DNA is due to the presence of several repetitive elements across the genome. Some of the common non-B DNA structures include Z-DNA, H-DNA (triplex DNA), cruciform DNA, G-quadruplexes and RNA: DNA hybrid (R-loops).

Over the past few decades G-quadruplex structures have gained tremendous importance owing to its role in physiology and pathology. Recently it has been shown that novel sequence motifs, called GNG or bulges can fold into G-quadruplexes, thus increasing the propensity of such structures genome-wide. Neurological diseases, psychiatric diseases and genomic disorders (due to deletions, translocations, duplications and inversions) are some of the consequences of non-B DNA structures in the human genome.

Inadvertent genomic rearrangements in human can lead to different diseases including cancer. Immediate consequence of genomic rearrangement includes structural alteration of genome through joining of distant sequences. t(8;14) translocation is the hallmark of Burkitt's lymphoma, which results in deregulation of *c-MYC* gene that may contribute to oncogenic transformation. In the present study, we delineate the causes of fragility within the *c-MYC* gene. In order to do this, breakpoints at the *c-MYC* locus from Burkitt's lymphoma patient sequences reported in database were plotted and analysed. Interestingly, unlike many other

translocations, breakpoints at *c-MYC* locus were widespread, except for a cluster of breakpoints downstream to promoter 2 (P2).

Previous studies indicate that the translocation breakpoint clusters often correlate with formation of non-B DNA structures. The entire breakpoint cluster downstream of P2 was divided into Region 1, Region 2 and Region 3. Interestingly, *in silico* analysis of the breakpoint clusters revealed no evidence for predictive classic non-B DNA motifs in Region 2; whereas Region 1 harboured a G-quadruplex motif on the template strand and Region 3 had two short inverted repeats. Intriguingly, as the nontemplate strand of Region 2 was G skewed with a good number of AID binding motifs, we tested the MYC breakpoint Region 2 for its potential to form R-loop due to binding of nascent RNA to template DNA. Our results showed that MYC Region 2 can form RNA-DNA hybrid in a transcription dependent manner in physiological orientation. Observed structure was sensitive to RNase H. We showed Region 2 hindered action of Dpn I upon transcription confirming formation of R-loop structure. Owing to single strandedness, Region 2 R-loop was shown to be sensitive to P1 nuclease as opposed to the untranscribed control. The single strandedness of the Region 2 R-loop was characterized at a single molecule level through bisulfite modification assay. The assay corroborated formation of R-loop along with providing snapshots of various length R-loops formed upon Region 2 transcription. Besides, various biophysical and biochemical assays showed the complementary region (template strand) to be single-stranded in stretches, upon transcription. Length of RNA within the R-loop was within a range of 75 to 250 nt. To delineate the mechanism of R-loop formation we tested the sensitivity of R-loop formation to RNase A during and post transcription; and found that R-loop formation was abrogated in presence of RNase A during transcription suggesting that R-loop formation followed a “thread back model”.

Intriguingly we observed that two short regions of the template strand exhibited high degree of single strandedness. To investigate the reason for such unusual single strandedness, oligonucleotides spanning the region was designed and subjected for CD and EMSA studies. EMSA showed robust intramolecular G-quadruplex structure formation in presence

of KCl, whereas CD confirmed that both regions formed parallel G-quadruplexes. We also showed the precise involvement of guanines in structure formation through DMS protection assay. Further, the region of interest was cloned into appropriate vectors and primer extension assays were performed in presence of G-quadruplex stabilizing agents like TMPyP4 and KCl. Increasing concentration of these stabilizing agents enhanced the formation of G-quadruplexes in a double stranded context, which hindered polymerase progression. Since these G-quadruplex structures utilized sequences which are deviant to the consensus of G-quadruplex motifs, non-B DNA predicting tools were unable to score them. On closer analysis of the sequences we found that, these G-quadruplexes involve duplex hairpin and GNG motifs during structure formation. Besides, both the G-quadruplexes were highly thermostable and were able to fold back upon renaturation.

Till recently, it has been believed that G-quadruplex structures are formed using a minimum of four, 3 guanine tracts, with connecting loops ranging from one to seven. Recent studies have reported deviation from this general convention. One such deviation is the involvement of bulges in the guanine tracts. In the present study, guanines along with GNG motifs have been extensively studied using recently reported HOX11 breakpoint fragile region I as a model template. By strategic mutagenesis approach we show that the core elements of a G-quadruplex are not equally important in structure formation when flanked by GNG motifs. Importantly, the positioning and number of GNG/GNGNG can dictate the formation of G-quadruplexes. In addition to HOX11 fragile region, GNG motifs of HIF1-alpha can fold into intramolecular G-quartet. However, GNG motifs in mutant VEGF sequence could not participate in structure formation, suggesting that the usage of GNG is context dependent. Importantly, we show that when two stretches of guanines are flanked by two independent GNG motifs in a naturally occurring sequence (SHOX), it can fold into an intramolecular G-quadruplex. Interestingly, intra molecular GNG G-quadruplexes were able to fold back after complete denaturation of the oligonucleotides. Besides one of the intra molecular GNG G-quadruplexes was purified and confirmed for parallel conformation. Finally, we show the specific binding of G-quadruplex binding protein, Nucleolin and G-quadruplex antibody BG4

to SHOX G-quadruplex through EMSA studies. Thus, the study provides novel insights into the role of GNG motifs in G-quadruplex structure formation, which may have both physiological and pathological implications.

In conclusion, we show formation of transcription dependent R-loop and G-quadruplex structures at the *c-MYC* gene locus in a mutually exclusive manner. The data presented here, in conjunction with studies from other laboratories suggests that these structures could impart fragility within the *c-MYC* gene locus during t(8;14) translocation. Besides, we characterised unusual G-quadruplexes harbouring GNG motifs. We find that positioning and number of GNG can dictate the formation of G-quadruplexes and is context dependent.